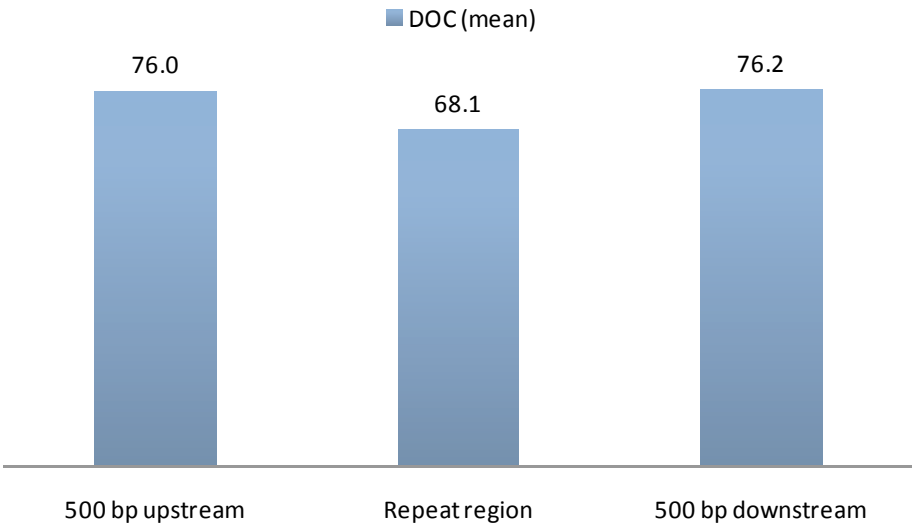
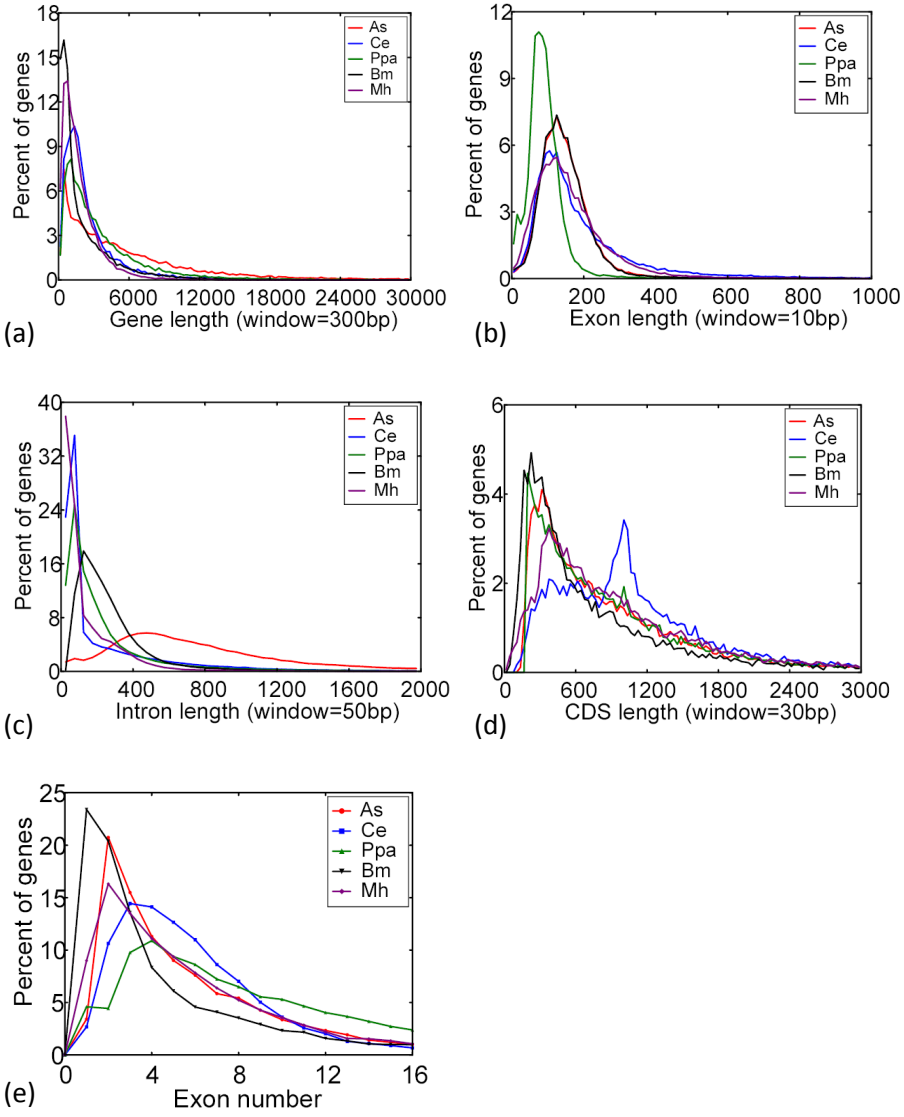
doi:10.1038/nature10553

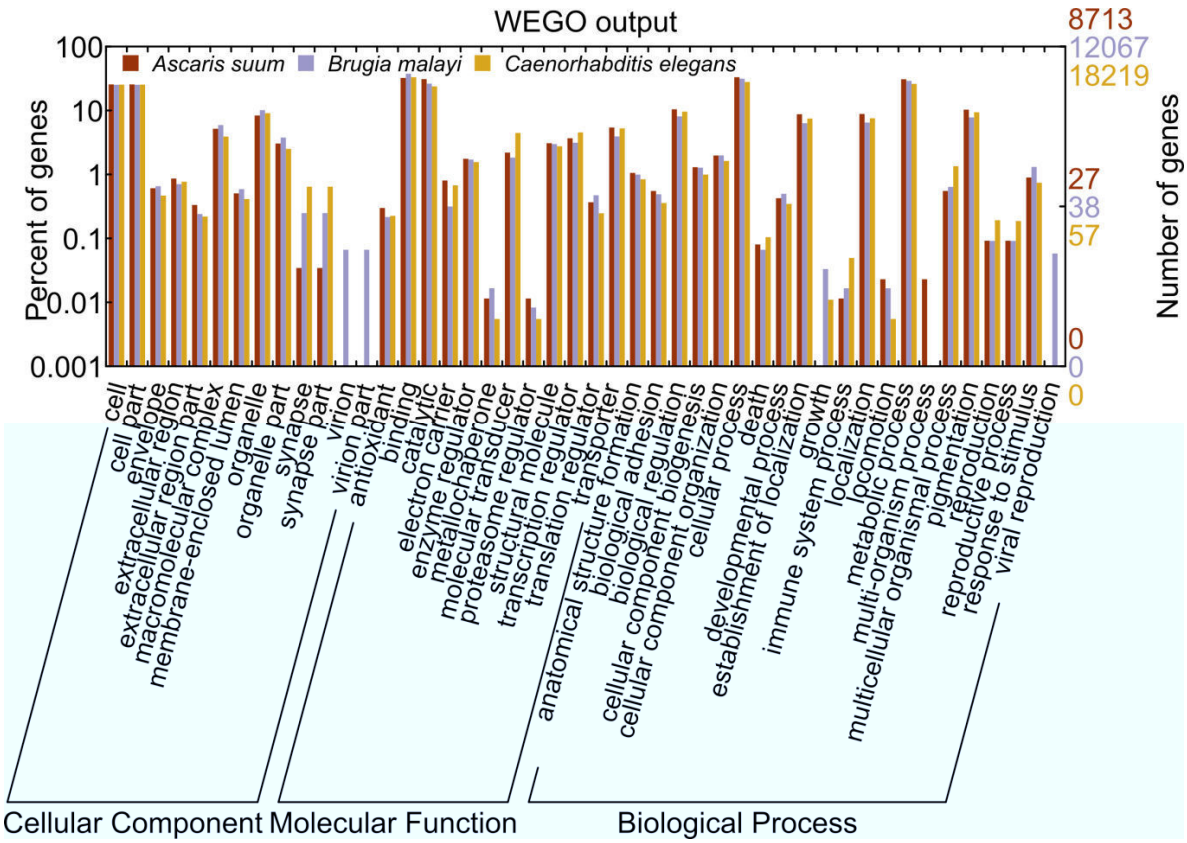
Supplementary Figures



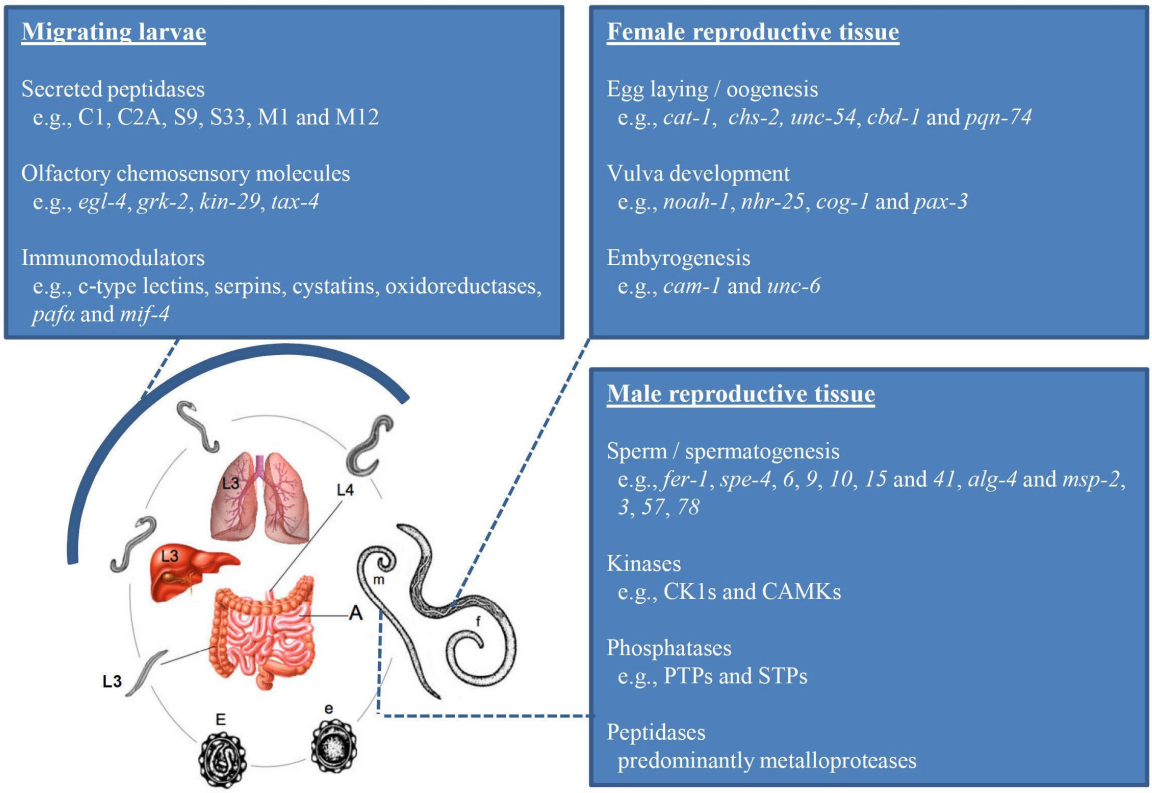
Supplementary Figure 1 | Comparative assessment of depth of coverage (DOC) of reads by mapping to the repetitive and flanking, non-repetitive regions of the *Ascaris suum* genome.



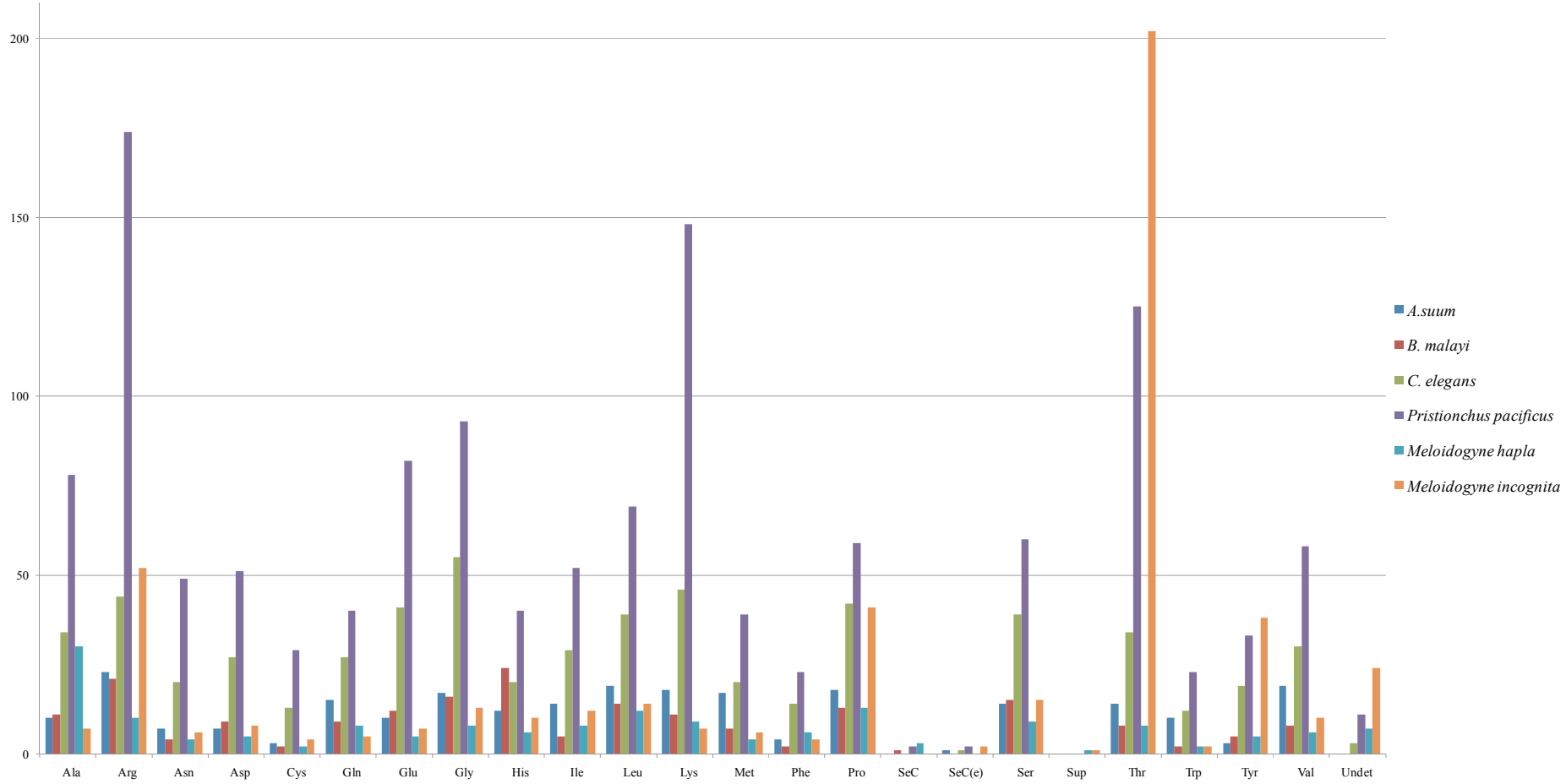
Supplementary Figure 2 | Proportional distribution of the major gene characteristics of key species of Nematoda. *Ascaris suum* (As), *Caenorhabditis elegans*[2] WS210 (Ce), *Pristionchus pacificus*[3] WS222 (Ppa), *Brugia malayi*[3] WS222 (Bm), *Meloidogyne hapla*[4] WS222 (Mh). (a) Total gene length, (b) exon length, (c) intron length, (d) coding sequence (CDS) length, and (e) exons per gene.



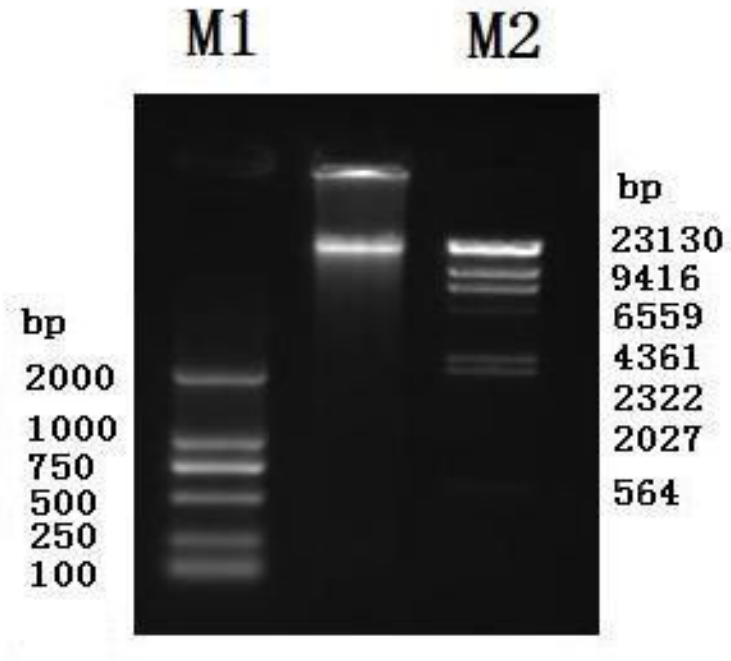
Supplementary Figure 3 | Comparative distribution of GOSlim[7] classifications of all predicted protein-coding genes for *Ascaris suum*, *Brugia malayi* and *Caenorhabditis elegans*. The x-axis depicts all level 2 GOSlim terms inferred using the program InterProscan[8] for each species. The y-axis is $\log_{10}$-transformed with the left y-axis, representing the proportion of genes (%) assigned to each level 2 GOSlim classification, and the right axis represents absolute numbers of genes normalized relative to the total number of genes with a GO classification in each genome (i.e., 8,713 for *A. suum*, 12,067 for *B. malayi* and 18,219 for *C. elegans*) at an e-value cut-off of 10^{-5}. This figure was generated using the online tool WEGO (available via http://wego.genomics.org.cn/cgi-bin/wego/index.pl).



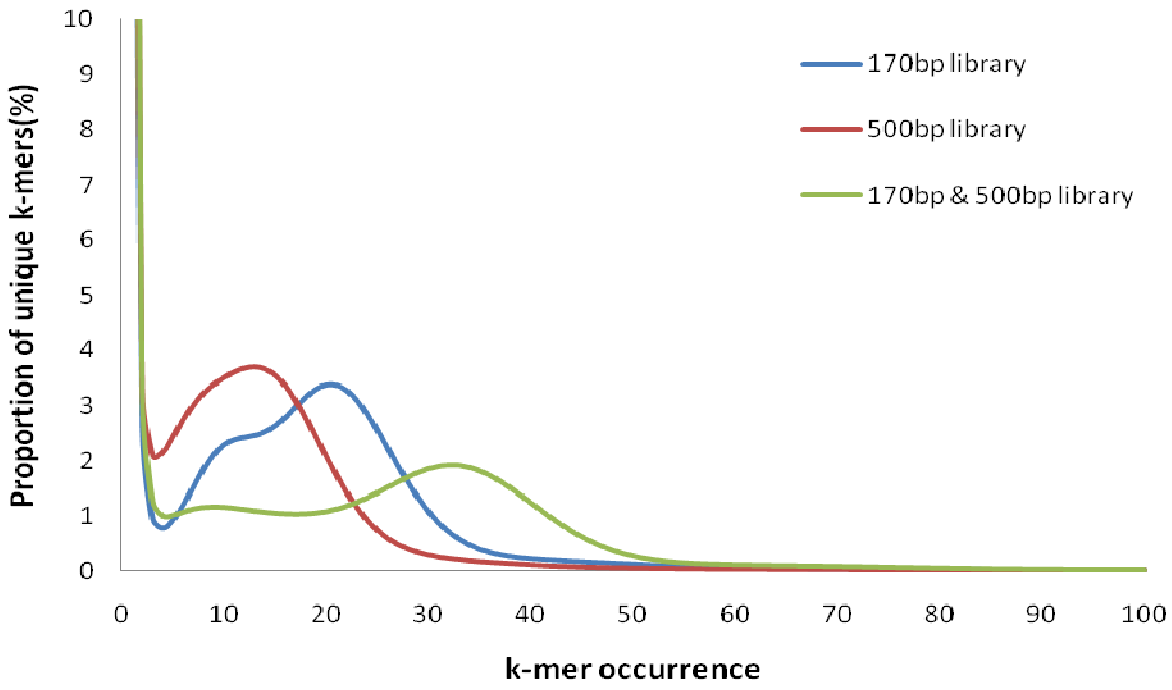
Supplementary Figure 4 | Key changes in gene transcription associated with various phases of the *Ascaris suum* life cycle. Schematic indicating gene categories for which there is increased transcription during hepatopulmonary migration (larval stage) or within male or female reproductive tissues (adult stage). Abbreviations (clockwise): e, egg; E, larvated egg; L3, third-stage larva; L4, fourth-stage larva; A, adult; F, adult female, M, adult male; *pafα*, platelet anti-inflammatory α; *mif-4*, macrophage inhibition factor 4; CK1, casein kinase; PTP, protein tyrosine phosphatase; STP, serine/threonine phosphatases.



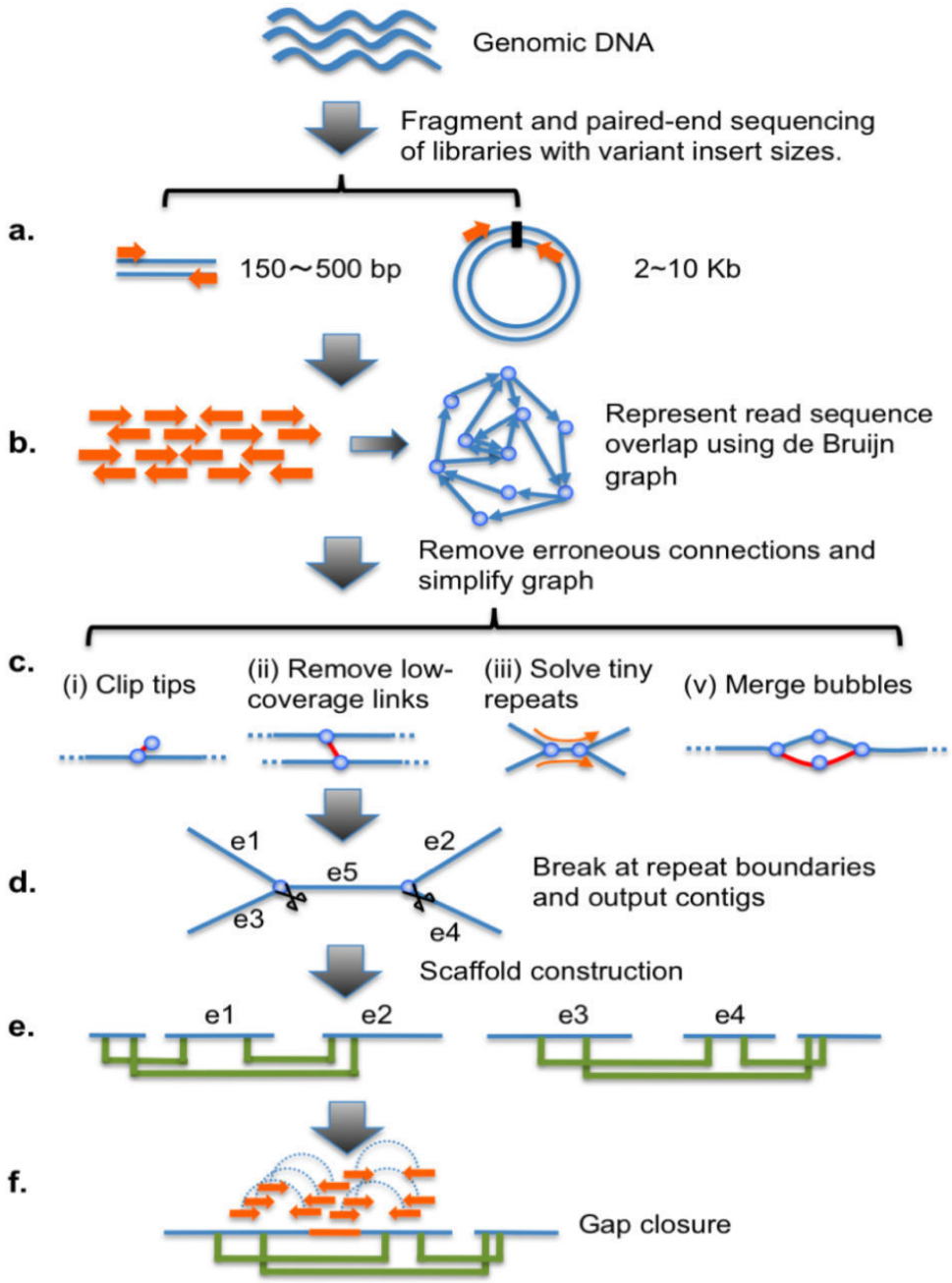
Supplementary Figure 5 | Comparison of copy number of transfer RNA genes for key species of nematodes. Data based on tRNA-ScanSE predictions for *Ascaris suum*, *Caenorhabditis elegans*[2], *Brugia malayi*[3], *Pristionchus pacificus*[5], *Meloidygne hapla*[4] and *Meloidogyne incognita*[6].



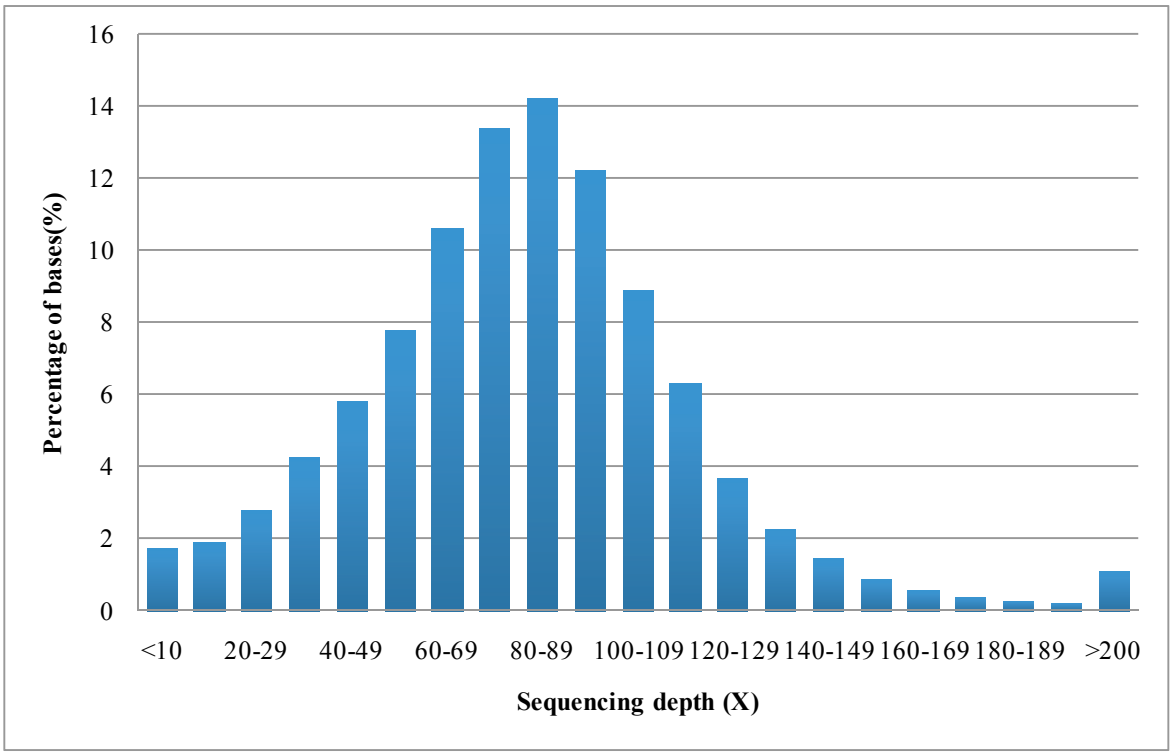
Supplementary Figure 6 | High molecular weight DNA used for mate-paired library construction (2000, 5000 and 10,000 bp -insert libraries). Whole genome amplification (WGA), employing the REPLI-g Midi Kit (Qiagen), was used to produce (from 200 ng of genomic template) the amount of DNA required for the construction of the 2 kb, 5 kb and 10 kb libraries. Lanes M1 (DL2000 DNA Marker, Tiangen) and M2 (λ-Hind III Marker, Takara) are molecular weight markers.



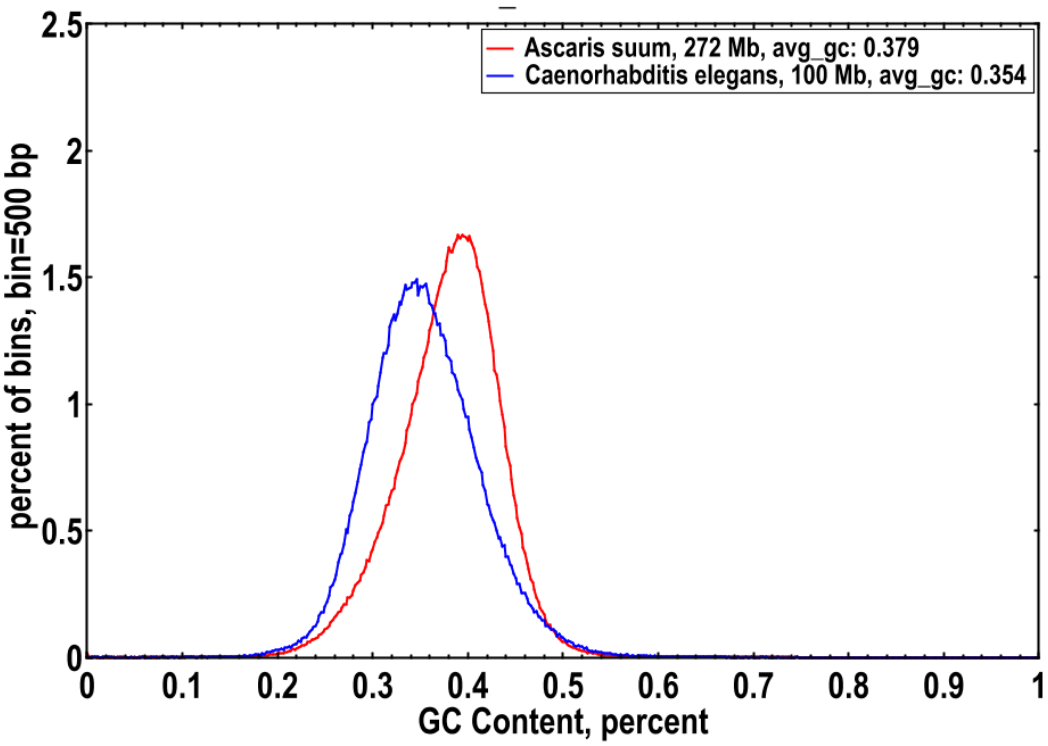
Supplementary Figure 7 | The frequency of unique 17-mers within *Ascaris suum* genomic DNA using 170, 170 and 500 or 170, 500 and 800 bp -insert libraries.



Supplementary Figure 8 | A schematic representation of the genome assembly strategy conducted using SOAPdenovo (image reproduced from[1]). (a) Short-insert libraries (170 bp, 500 bp and 800 bp) were constructed from linearised DNA fragments; large-insert libraries (2 kb and 10 kb) were generated from circularised, mate-paired constructs; (b) following sequencing on the Illumina HiSeq platform, all reads were used to construct De Bruijn graphs representing all overlapping regions between and among the reads; (c) artefactual and or erroneous connections within the De Bruijn graph (i.e., single artefact reads, low coverage areas, small repeats and redundant "bubbles" in the De Bruijn construct) were removed; (d) larger repetitive regions of the De Bruijn graph assembly were further corrected and all unambiguous connections among reads were used to generate the resulting sequence contigs; (e) using paired-end data from each sequence library, these contigs were connected into larger scaffold sequences; (f) where possible, gaps (i.e., N's) were filled in using the paired-read data to generate the final assembly.



Supplementary Figure 9 | Distribution of sequence depth of the assembled *Ascaris suum* genome. The Illumina reads were aligned on to the assembled genome sequence using SOAP-aligner, with a maximum threshold of 5 mismatches per read.



Supplementary Figure 10 | GC-content distribution for the genomes of *Ascaris suum* and *Caenorhabditis elegans*. Data generated using 500 bp sliding windows using a 250 bp overlap. The x-axis is GC content (%), and the y-axis is the proportion of the windows (=bins) corresponding to each GC measurement.

References

1 Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**, 265-272 (2010).
2 CSC. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).
3 Ghedin, E. *et al.* Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756-1760 (2007).
4 Opperman, C. H. *et al.* Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism. *Proc Natl Acad Sci U S A* **105**, 14802-14807 (2008).
5 Dieterich, C. *et al.* The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat Genet* **40**, 1193-1198 (2008).
6 Abad, P. *et al.* Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nat Biotechnol* **26**, 909-915 (2008).
7 Camon, E. *et al.* The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res* **13**, 662-672 (2003).
8 Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, W116-120 (2005).

Supplementary Tables

Supplementary Table 1. Prediction of repetitive regions in the *Ascaris suum* genome

Type	Repeat size	% of genome
RepeatProteinmask	1,198,412	0.44
Repeatmasker	1,714,809	0.63
TRF	5,747,739	2.11
De novo [a]	10,856,746	3.98
Total (overlapping)	17,201,326	6.31
Total (non-overlapping)	10,542,487	4.38

[a] Includes data generated by Piler, LTR-Finder and RepeatScout.

Supplementary Table 2. Repeats derived from *de novo* and homology-based predictions

	Repbase TEs		**TE proteins**		***de novo***		**Combined** [c]	
Repeat-type	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome
DNA	935,439	0.34	120,504	0.04	1,955,159	0.72	2,959,712	1.09
LINE	436,363	0.16	684,849	0.25	4,489,808	1.65	4,885,744	1.79
SINE	11,843	0.00	0	0.00	1,328	0.00	12,928	0.00
LTR	451,801	0.17	393,059	0.14	3,404,975	1.25	3,890,554	1.43
Other[a]	1,145	0.00	0	0.00	103,915	0.04	105,060	0.04
Unknown[b]	29,689	0.01	0	0.00	1,201,096	0.44	1,219,901	0.45
Total	1,714,809	0.63	1,198,412	0.44	10,542,487	3.86	11,938,602	4.38

[a] "Other" refers to the repeats classified by RepeatMasker, which are not included in the other groups;
[b] "Unknown" refers to the predicted repeats that cannot be classified by RepeatMasker.
[c] "Combined" represents the non-redundant consensus of all repeat prediction/classification methods employed.

Supplementary Table 3. Classification of repeats in the *Ascaris suum* genome

Type[a]	Number	% of total
Retrotransposons	424	68.2
LINE/Cr1	29	4.7
LINE/L1	28	4.5
LINE/RTE-RTE	21	3.4
LINE/Penelope	14	2.3
LINE/R2	5	0.8
LINE/L2	4	0.6
LINE/Jockey	3	0.5
LINE/Dong-R4	2	0.3
LINE/LOA	2	0.3
LINE/RTE-BovB	2	0.3
LINE/Telomeric	2	0.3
LINE (others)	6	1
LTR/Gypsy	97	15.6
LTR/Pao	85	13.7
LTR/Copia	60	9.7
LTR/ERV1	19	3.1
LTR/ERVK	9	1.5
LTR/ERVL	3	0.5
LTR/Ngaro	3	0.5
LTR(other)	7	1.1
SINE/tRNA	9	1.5
SINE (others)	1	0.2
Retrotransposons (others)	14	2.3
DNA transposons	91	14.6
DNA/MuDR	12	1.9
DNA/En-Spm	9	1.5
DNA/Merlin	8	1.3
DNA/Maverick	6	1
DNA/hAT-Charlie	6	1
DNA/TcMar-Tc4	5	0.8
DNA/TcMar-Tc1	5	0.8
DNA/Harbinger	4	0.6
DNA transposons (others)	36	5.8
Satellite	10	1.6
Unknown	96	15.4
Total	622	100

[a]All repeat types were assigned according to homology to the Repbase database (http://www.girinst.org/repbase).

Supplementary Table 4. Statistics of predicted protein-coding genes in the *Ascaris suum* genome

Gene-set predicted from/using	No. of genes inferred	Mean gene length (bp)	Mean coding sequence length (bp)	Mean GC content (%)	Mean no. of exons per gene	Mean exon length (bp)	Mean intron length (bp)
RNA-seq	64,569	2,203	532.8	41.1%	2.5	216.3	1,141
Augustus	37,801	2,292	523.9	44.6%	3.5	149.8	708
GlimmerHMM	48,471	2,714	499.3	44.7%	3.3	152.8	976
SNAP	51,547	948	393.0	44.6%	2.7	145.7	327
Human/*C. elegans*	9,155	7,017	1,051.3	45.6%	7	149.7	990
Combined (Glean)	18,542	6,536	983.2	45.3%	6.4	153	1,081

Supplementary Table 5. Assessment of the quality of the *Ascaris suum* genome assembly using RNA-seq (expressed sequence tag, EST) data

EST contigs	Total Number	Total number mapped	% that mapped	> 50% of sequence covered by one scaffold		> 90% of sequence covered by one scaffold	
				Number	%	Number	%
All	199,283	189,821	95.25	189,816	95.25	183,811	92.24
> 200 bp	129,506	121,767	94.02	121,762	94.02	119,810	92.51
> 500 bp	39,279	38,425	97.83	38,424	97.82	38,188	97.22
> 1000 bp	14,900	14,263	95.72	14,263	95.72	14,191	95.24

For the mapping of EST, contigs are defined as continuous strings of sequence with N's. So defined, all EST contigs were aligned to the *A. suum* assembly using BLAT.

Supplementary Table 6 | Key functional and druggable protein families, predicted to be encoded in the *Ascaris suum* genome based on BLAST comparisons (10^{-5}) to specialized curated databases (see Supplementary Methods)

Family	BLASTp (10^{-5})
Peptidases	456
Metallo-protease	184
Serine protease	132
Cysteine protease	90
Threonine protease	22
Aspartic protease	18
Kinases	609
TK	94
CK1	83
CMGC	67
CAMK	54
AGC	34
STE	34
RGC	15
TKL	20
Other	72
Atypical	9
Phosphatases	257
Protein-tyrosine	68
Receptor protein tyrosine	17
Serine-threonine	64
Dual specificity	39
GTPases	169
Receptors	649
GPCRs	279
Transporters	1797
Channels/pores	477
Porters	462
P-P-bond-hydrolysis driven transporters	382
Auxillary transport proteins	104

Supplementary Table 7 | Summary of the *Ascaris suum* excretory/secretory proteins, predicted to play a key role in immuno-modulation, -regulation and/or -evasion in the host animal

Role	Number of genes (%)
Immunogen	301 (78.4)
O-linked glycan	300
Omega-1 ribonuclease	1
Immunogen/inhibitor	17 (4.4)
SCP/TAPs (VAL)[a]	17
Inhibitor/immunomodulator	26 (6.8)
Aminopeptidase	2
Cystatin	2
Serpin	11
SmSP1[b]	1
Galectin	8
Calreticulin	2
Mimicry	18 (4.7)
C-type lectin	7
Lectin	6
MIF[c]	2
TGF-β like	2
Oxidoreductase	14 (3.6)
Glutathione peroxidase	5
Peroxidase	1
Superoxide dismutase	3
Thioredoxin peroxidase	5
Anti-inflammatory	5 (1.3)
PAFA[d]	5
Others	3 (0.8)
ALT1,2[e]	3

[a]SCP/TAPS (VAL), Sperm Coating Protein/Tpx-1/Ag5/PR-1/Sc7 (venom allergen-like) proteins, [b]SmSP1, serine protease inhibitor (serpin) 1 encode in *Schistosoma mansoni* (human blood fluke), [c]MIF, Macrophage initiation factor 4 mimic, [d]PAFA, platlet anti-inflammatory α, [e]ALT1, 2, abundant larval transcripts 1 and 2, encoded in *Brugia malayi*.

Supplementary Data Descriptions

Supplementary Data 1 | Synteny comparisons. This data spreadsheet provides a summary of the pairwise comparisons of overall chromosomal (sheet 1), inter- (sheet 2) and intrachromosomal gene (sheet 3) synteny between *Ascaris suum* or *Brugia malayi* and *Caenorhabditis elegans*. Reciprocal BLASTp analyses were conducted to identify one-to-one gene orthologues (e-value threshold: 10^{-5}). These orthologous were then mapped to respective scaffolds of the *A. suum* or *B. malayi* assembly; those that mapped to large scaffolds (i.e > 1 Mb) were aligned to *C. elegans* chromosomes. Using this approach, we quantified the numbers of one-to-one orthologues for each large *A. suum* or *B. malayi* scaffold that mapped to each of the *C. elegans* chromosomes and assessed the relative 'patchiness' of the distribution of these genes by Shannon Diversity Index, reasoning that the more frequent the rate of interchromosomal re-arrangements, the more homogenized the distribution of genes would become when comparing large scaffolds to each of the *C. elegans* chromosomes.

Supplementary Data 2 | *A. suum*, *B. malayi* and *C. elegans* orthologues. This data spreadsheet presents the identity of and functional information for orthologous genes uniquely shared between *Ascaris suum* or *Brugia malayi* and *Caenorhabditis elegans*, respectively. Sheet 1 presents *C. elegans* orthologues found in *A. suum* and absent from the current assembly for *B. malayi* and including the *A. suum* gene-code (column 1), *C. elegans* orthologue and the e-value of the BLASTp comparison linking these genes (column 3). Sheet 2 represents a summary of the functional ontology (GO) of each *C. elegans* orthologues, and is presents as the *C. elegans* gene-code (column 1), the GO code (column 2), as well as, the category 1 (column 3) and category 2 (column 4) GO terms associated with each GO code. A summary of these data is also provide. Sheet 3 and Sheet 4 represent information for the *C. elegans* orthologous genes found in *B. malayi* and absent from the current assembly for *A. suum* and the corresponding function ontology (GO) of these genes and is formatted in the same manner as sheets 1 and 2.

Supplementary Data 3 | *A. suum* vs *C. elegans* KEGG pathways. This data spreadsheet presents a comparison of the number of genes mapping to each biological pathway defined by the Kyoto Encyclopaedia of Genes and Genomes (KEGG) and encoded in the *A. suum* or *C. elegans* genome. The data represent each KEGG orthology (ko) code determined for each gene-set (column 1), relating to the conserved KEGG pathway(s) to which this orthologue belongs (column 2). In addition, we present the total number of orthologues detected for each KO code/pathway category for *A. suum* (column 3) and *C. elegans* (column 4), and the numerical difference in these values (column 5). These data are summarized graphically in a bar chart, showing the comparative pathways present in the *A. suum* (blue) and *C. elegans* (red) genomes, comparing data from columns 3 and 4.

Supplementary Data 4 | Key protein families. The data spreadsheet provides a gene-by-gene classification (based on BLAST homology to key reference databases) for each of the

major protein classes encoded in the *Ascaris suum* genome and including GTPases (sheet 1), G-protein coupled receptors (sheet 2), kinases (sheet 3), peptidases (sheet 4), phosphatases (sheet 5) and transporters and channel proteins (sheet 6). Each gene is represented by a gene-code which matches to each annotated coding domain predicted from the *A. suum* genome (column 1), the nearest BLAST homologue for each gene (second to last column) and the predicted e-value associated with this match (last column). Where appropriate, each protein class is further sub-classified. A summary table is also presented for each protein class, and summarized graphically as a pie-chart.

Supplementary Data 5 | ES proteins. This data spreadsheet presents a gene-by-gene classification for the predicted secretome (i.e., excretory/secretory [ES] peptides) for *Ascaris suum*, represented by a gene-code that matches each annotated coding domain predicted from the *A. suum* genome (column 1), the nearest BLAST homologue for each gene (second to last column) and the predicted e-value associated with this match (last column). Because of their specific relevance in the biology of *Ascaris suum*, predicted peptidases and their inhibitors are characterized further, sub-classified as appropriate (sheets 2 and 3, respectively). A summary table and pie-chart are provided for each sheet.

Supplementary Data 6 | Immunomodulators. This data spreadsheet provides a gene-by-gene characterization of each *Ascaris suum* ES protein with close homology to a known immunomodulatory secreted protein derived from a species of helminth (i.e., a species of Nematoda, Digenea or Cestoda). Each gene is associated with a specific gene-code that matches to each annotated coding domain predicted for the *A. suum* genome (column 1), the nearest BLAST homologue (column 2), the protein class (column 3), the e-value associated with the homology match (column 4), the general immunomodulatory role of the peptide based on experimental data (column 5) and the specific immunomodulatory role for the peptide based on the literature (column 6). A summary of these data is provided in Supplementary Table 7.

Supplementary Data 7 | Larval RNA-seq heatmaps. This compressed (*.zip) data file contains heat-maps displaying the differential transcription in *Ascaris suum* during the hepatopulmonary migration phase. A detailed description of these files is included in the compressed folder.

Supplementary Data 8 | Adult RNA-seq heatmaps. This compressed (*.zip) datafile contains heat-maps displaying the differential transcription between adult *Ascaris suum* male and female muscle or reproductive tissue. A detailed description of these files is included in the compressed folder.

Supplementary Data 9 | Single nucleotide polymorphisms (SNPs). This data spreadsheet summarizes single nucleotide polymorphism (SNP) data for the multiple *Ascaris suum* individuals represented in the transcriptomic libraries sequenced here, relative to the genome assembly. Sheet 1 provides a detailed characterization of each SNP detected for each gene. Each gene is represented by its gene code, with each SNP location provided as a base-pair value relative to the length of the gene (column 2). In addition, the nucleotide sequence of the genome assembly at the putative SNP location is provided (column 3), as is the value of the SNP itself (column 4). To allow an evaluation of the effect of the SNP on the protein sequence, the region of the gene containing an SNP was translated in frame for the reference (i.e., genomic) sequence (column 5) or the transcriptomic data (column 6) and then summarized in column 7 as synonymous (i.e., with no amino acid change), non-synonymous (i.e, with an amino acid change), stop (i.e., introducing a stop codon) and ambiguous (i.e., represented by an IUPAC code for which one value results in a synonymous mutation, whereas the other results in a non-synonymous mutation). These SNP classifications are also summarized in a table and a pie chart. Sheet 2 provides a gene-by-gene summary of the number of SNPs and their effect on the encoded protein and is presented as a gene-code (column 1), followed by the number of ambiguous (column 2), non-synonymous (column 3), stop (column 4), synonymous (column 5) and total (column 6) SNPs. Sheet 3 provides a gene-by-gene ranking of variation based on the number of SNPs per thousand base-pairs (SNPK) for each gene for which at least 100 sequence reads were available in the transcriptomic data sets and is presented as a gene-code (column 1), followed by total reads (column 2), absolute number of SNPs (column 3), SNPK (column 4), gene length (column 5) and ordinal ranking (column 6). Sheet 4 presents available KEGG orthology data for the most variable 2.5% of these genes (i.e., the top 2.5%, based on ranking). Sheet 5 represents available KEGG orthology data for the most conserved 2.5% of these genes (i.e., the bottom 2.5%, based on ranking).

Supplementary Data 10 | Druggable *A. suum* genome. This data spreadsheet provides a detailed summary of essentiality and metabolic chokepoint-based prediction of the druggable *Ascaris suum* genome. Sheet 1 provides a gene-by-gene annotation of the *A. suum* genes with close homology to genes with adult lethal phenotypes based on knockout and/or knockdown data for *Caenorhabditis elegans* and *Drosophila melanogaster*. Sheet 2 provides a summary of the available information in the CHEMBL database for known inhibitors for proteins encoded by these genes, including the CHEMBLid (column 2) and corresponding UniProt code (column 3) and, where applicable, EC number (column 4) for the nearest homologous sequence in the CHEMLB database (column 5), as well as a list of known inhibitors (column 6), their synonyms based, for example, on commercially available pharmaceuticals (column 7) and their mechanism of activity, if known (column 8). Sheet 3 presents a detailed summary of all metabolic chokepoints associated with substrates uniquely consumed by an essential (i.e., adult lethal) gene product from sheet 1, providing the identity of the unique substrate (column 2), the EC number of the essential enzyme (column 3) and its identity (column 4), all

substrates (column 5) and products (column 6) associated with the reaction and the conserved pathway to which each of these enzymes belong (column 7). Sheet 4 provides similar information to that given for sheet 3, but presents enzymes associated with the generation of a unique product. Sheet 6 provides a summary, in the same format as sheets 4 and 5, for single-copy genes associated with these chokepoints.